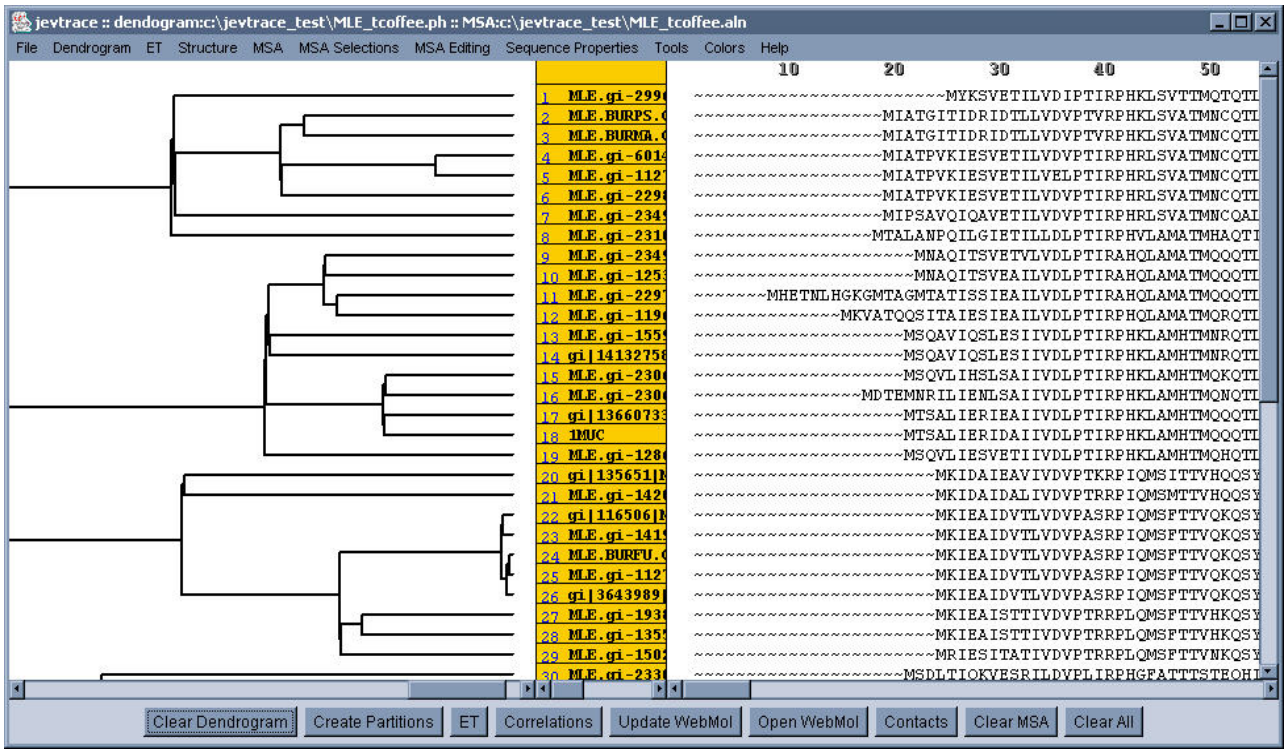
Jevtrace v3.12b



User Manual

Marcin Joachimiak

1/6/04

0. Overview

Jevtrace is a Java implementation of the Evolutionary Trace (ET) method and its variations. Jevtrace expands on the ET method, allowing detailed interactions with the underlying data as well as results of the analysis. The graphical interface allows protein family exploration in context of multiple sequence alignments, phylogeny and protein structures. Protein structure viewing and analysis is performed with WebMol, which is integrated and distributed with Jevtrace.

Jevtrace reference:

JEvTrace: refinement and variations of the evolutionary trace in JAVA.
Joachimiak MP, Cohen FE.
Genome Biol. 2002;3(12):RESEARCH0077. Epub 2002 Nov 26.

Selected Evolutionary Trace references:

Structural clusters of evolutionary trace residues are statistically significant and common in proteins.
Madabushi S, Yao H, Marsh M, Kristensen DM, Philippi A, Sowa ME, Lichtarge O.
J Mol Biol. 2002 Feb 8;316(1):139-54.

An evolutionary trace method defines binding surfaces common to protein families.
Lichtarge O, Bourne HR, Cohen FE.
J Mol Biol. 1996 Mar 29;257(2):342-58.

1. Installation

A. Jevtrace requires a working Java Runtime Environment (JRE). The Sun Microsystems J2E version 1.4.2 or the most current version is recommended (please do not use Java version 1.4.1 which has graphical problems).

Note: If you are using Windows or Macintosh it is likely that the operating system has a default Java setup. To check what version you are running, type 'java -version' on the command-line. Mac OS X stores previous versions of Java in the System directory.

B. Please fill out the Academic Software Licensing form and fax it to Erik Ellestad at 415-476-6515. You will be emailed a download link for the jevtrace package. For industrial licenses please contact Adelle Lohse. The software will sent to the provided email account.

C. Uncompress the Java package into a known directory.

D. Depending on OS and preference, proceed to setup or run Jevtrace.
The CLASSPATH modification allows for the easiest command line execution (**D.1)**.
Alternately one can use a shortcut, a script file such as a BAT (.bat) file on Windows, or setting the CLASSPATH through the command line '-cp' option (**D.2)**.

D.1. Add the directory from **C** to your system JAVA CLASSPATH. For example (on Windows):

set CLASSPATH c:\jevtrace_distr

For details on setting the CLASSPATH for all OS please refer to the web.
Sun Microsystems provides details on setting the CLASSPATH for Windows.

Windows 95 and 98 - Examine autoexec.bat for the set command, append Jevtrace directory to the 'set CLASSPATH' line.

Other (Windows NT, Windows 2000, ...) -The CLASSPATH environment variable can be set using the System utility in the Control Panel.

AND/OR D.2. If you do not feel like setting the CLASSPATH variable (or you need to start right away), you can use the –cp option to start Java, listing the relevant directories separated by a semicolon. Simply type the following to start Jevtrace:

java –cp c:\where\is\jevtrace\package\ jevtrace.jevtrace

Note: please do not add any other paths to the ‘-cp’ option.

2. The Input Data

A. Multiple Sequence Alignments (MSA).

Jevtrace is compatible with the CLUSTALW (.aln), FASTA and PILEUP (.msf), sequence alignment formats.

Currently the t_coffee multiple alignment program is suggested for accurate but slow alignments (especially cases of remote similarity), CLUSTALW/X is recommended for somewhat less accurate but much faster alignments. MAFFT is a newer multiple alignment algorithm that is even faster than CLUSTALW/X, however, no tree results are provided.

B. Phylogeny and trees

Jevtrace is compatible with the Newick (.ph, .dnd) and GCG (.figure) tree formats.

The PHYLIP and PAUP* software packages are suggested for phylogenetic tree reconstruction from protein sequence data. Some reliable and fast Neighbor-Joining methods for large data sets are FASTME and QuickTree. The methods mentioned in A will provide the input multiple alignment data for the phylogenetic analysis.

In many cases trees calculated from sequence similarity (such as the the .dnd output of the CLUSTALW progressive multiple alignment) can provide reasonable approximations of evolutionary relationships. However, whenever possible evolutionary data based on phylogenetic models is desired.

C. Protein Structure

PDB format files are compatible with Jevtrace and WebMol.

References:

Thompson JD, Higgins DG, Gibson TJ. Related Articles,
CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice.
Nucleic Acids Res. 1994 Nov 11;22(22):4673-80.

Notredame C, Higgins DG, Heringa J. Related Articles,
T-Coffee: A novel method for fast and accurate multiple sequence alignment.
J Mol Biol. 2000 Sep 8;302(1):205-17.

Katoh K, Misawa K, Kuma K, Miyata T. Related Articles,
MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform.
Nucleic Acids Res. 2002 Jul 15;30(14):3059-66.

FASTME:
Desper R, Gascuel O. Related Articles,
Theoretical Foundation of the Balanced Minimum Evolution Method of Phylogenetic Inference and Its Relationship to Weighted Least-Squares Tree Fitting.
Mol Biol Evol. 2003 Dec 23.

Howe K, Bateman A, Durbin R. Related Articles,
QuickTree: building huge Neighbour-Joining trees of protein sequences.
Bioinformatics. 2002 Nov;18(11):1546-7.

3. Starting Jevtrace

Jevtrace allows interaction with the following types of data:

	Sequence alignment	Phylogeny	Protein Structure
Sequence alignment	Yes[1]	Yes	Yes
Phylogeny	Yes	Yes[2]	Indirectly[3]
Protein Structure	Yes	Indirectly[3]	Yes[4]

[1] Crosstalk between different sequence alignments is achieved using a function to translate results (including graphical selections) from one alignment to another. At least one common identical sequence has to be present in both alignments.

[2] Crosstalk between different phylogenies or trees is currently not implemented in Jevtrace. Please refer to phylogenetic reconstruction packages like PAUP* or PHYLIP for methods to compare and analyze multiple trees.

[3] Crosstalk between phylogenies and structures (and vice versa) is achieved indirectly through the sequence alignment data and WebMol.

[4] Protein structure viewing and most analysis are delegated to WebMol.

The following are the command line startup options for Jevtrace:

java jevtrace.jevtrace

java jevtrace.jevtrace <session file>

java jevtrace.jevtrace <alignment file>

java jevtrace.jevtrace <tree file>

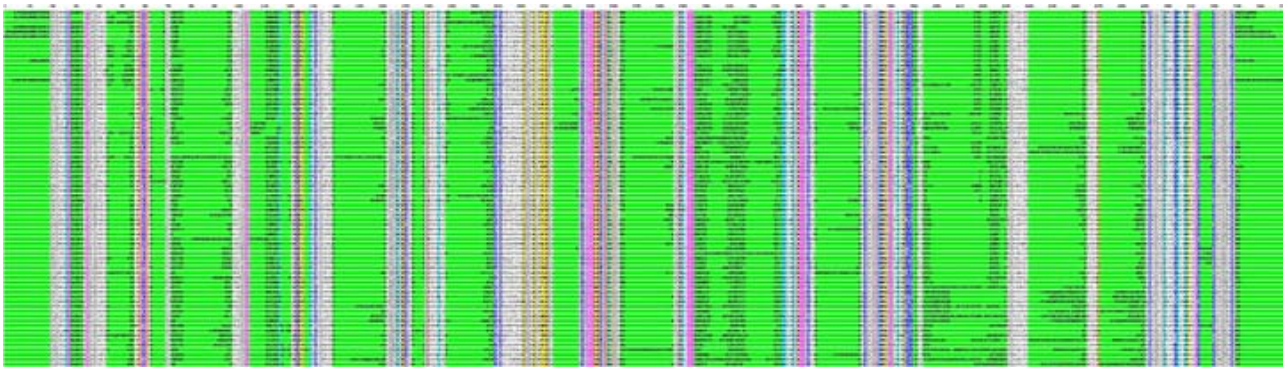
java jevtrace.jevtrace <tree file> <alignment file>

java jevtrace.jevtrace <tree file> <alignment file> <pdb file>

4. Talking to the Multiple Alignments

Jevtrace can be run with sequence data alone. The sequences can be in Fasta format (aligned or unaligned) or in the ALN alignment format used by programs such as ClustalW/ClustalX and t_coffee. The sequence-only mode of Jevtrace is especially useful when the set of sequences is large, when one is trying to prune the sequence set (for example removing redundant or spurious hits from a database search) or when phylogenetic and structure data are not available. Jevtrace allows an array of interactions with the sequence data, including a number of editing features. Details of editing functionalities can be found in **section 7**.

One example of Jevtrace sequence analysis functionality, is the ability to generate colored images of entire alignments as a single graphic ('**MSA**' menu '**Export MSA graphics**'). For example, the following graphic shows columns colored by percent sequence identity (using the color scale in **section 4.1**), and with columns containing gapped characters in green.



These images are of high resolution, such that any region can be zoomed in and viewed as legible characters.

4.1 Selecting amino acids, alignment columns and sequences

Selections in the sequence alignment (columns or individual characters) or the sequence names are the basis for performing editing operations as well as mapping sequence information to protein structures.

The Jevtrace coloring scheme consists of eight colors, which can be uniquely displayed and shaded on graphical representations of protein structure. This coloring scheme is also used as a color scale for scores generated by computations.

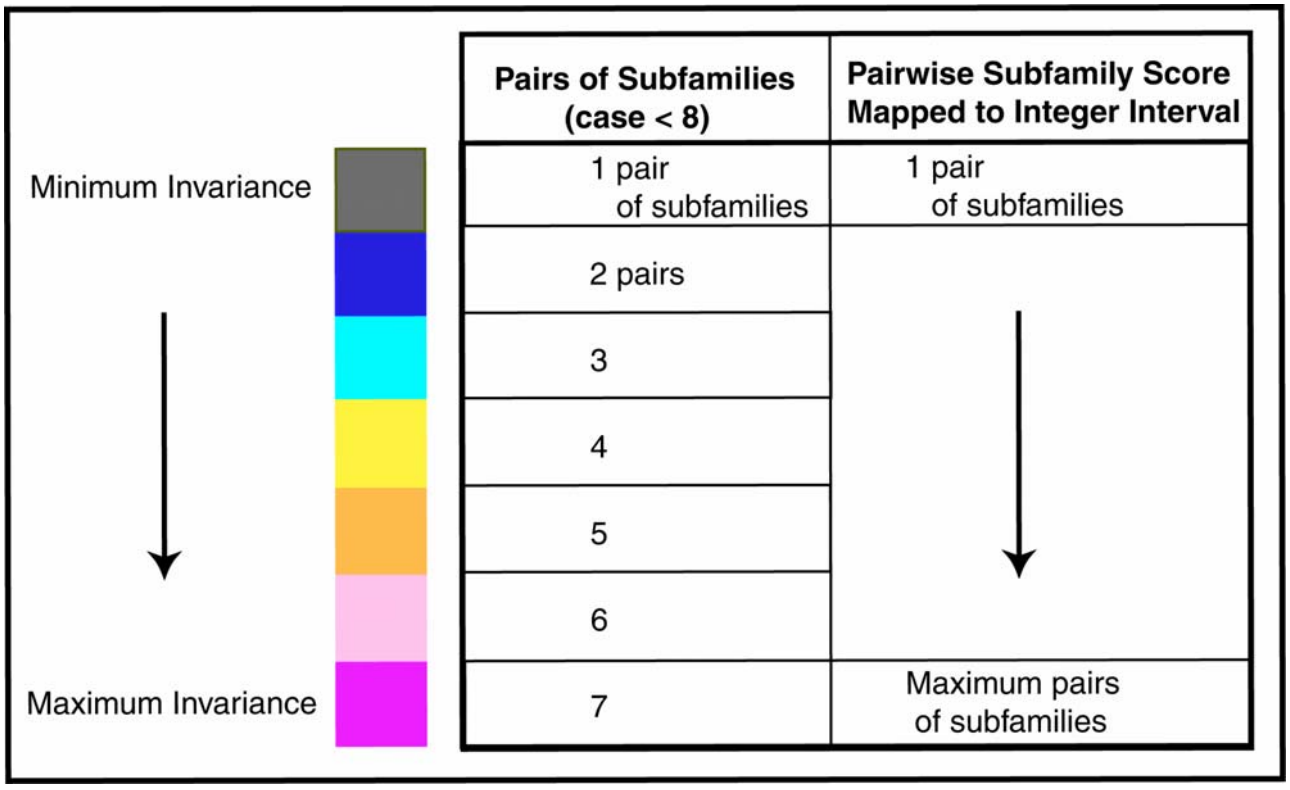


Figure 1

The color scale used in Jevtrace. The scale is used both to show cardinality (e.g. the left to right order of phylogenetic partitions) but also scores generated by computation.

Mouse clicking on any character, will select that character. Clicking again will deselect it.

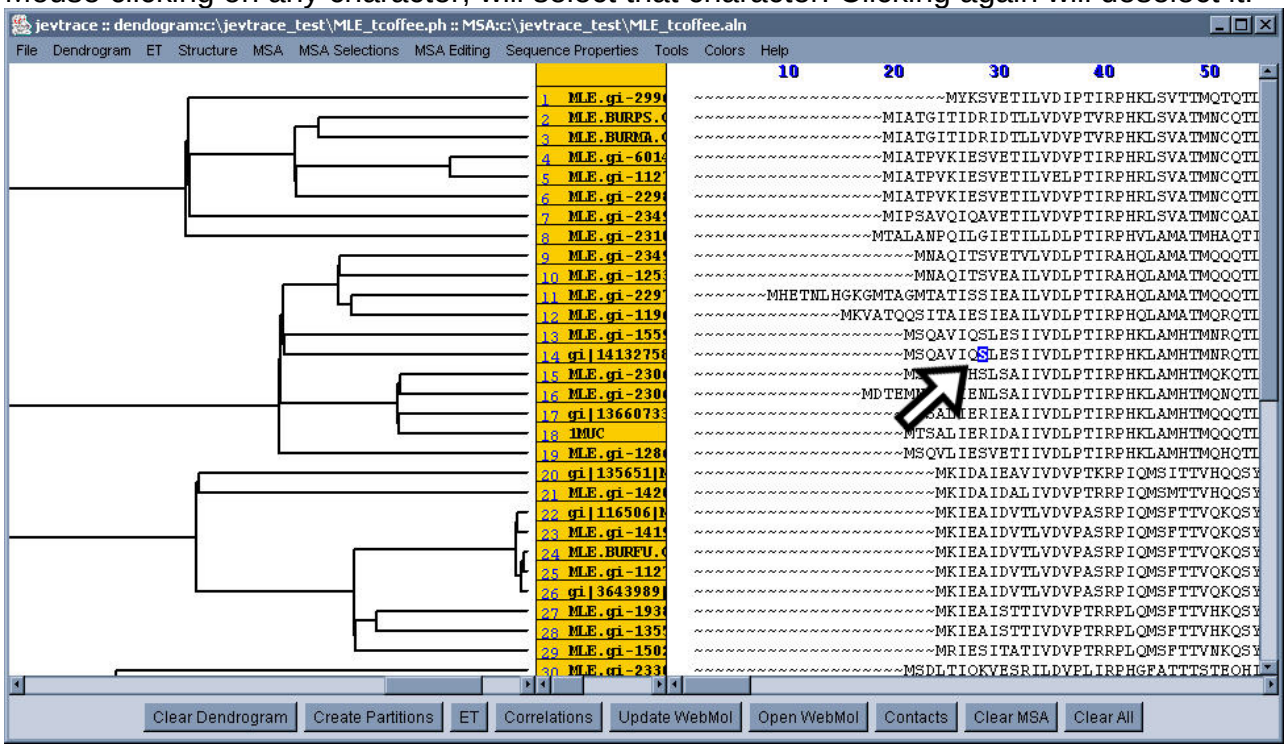


Figure 2
Mouse clicking on any character, will select that character (or deselect if already selected).

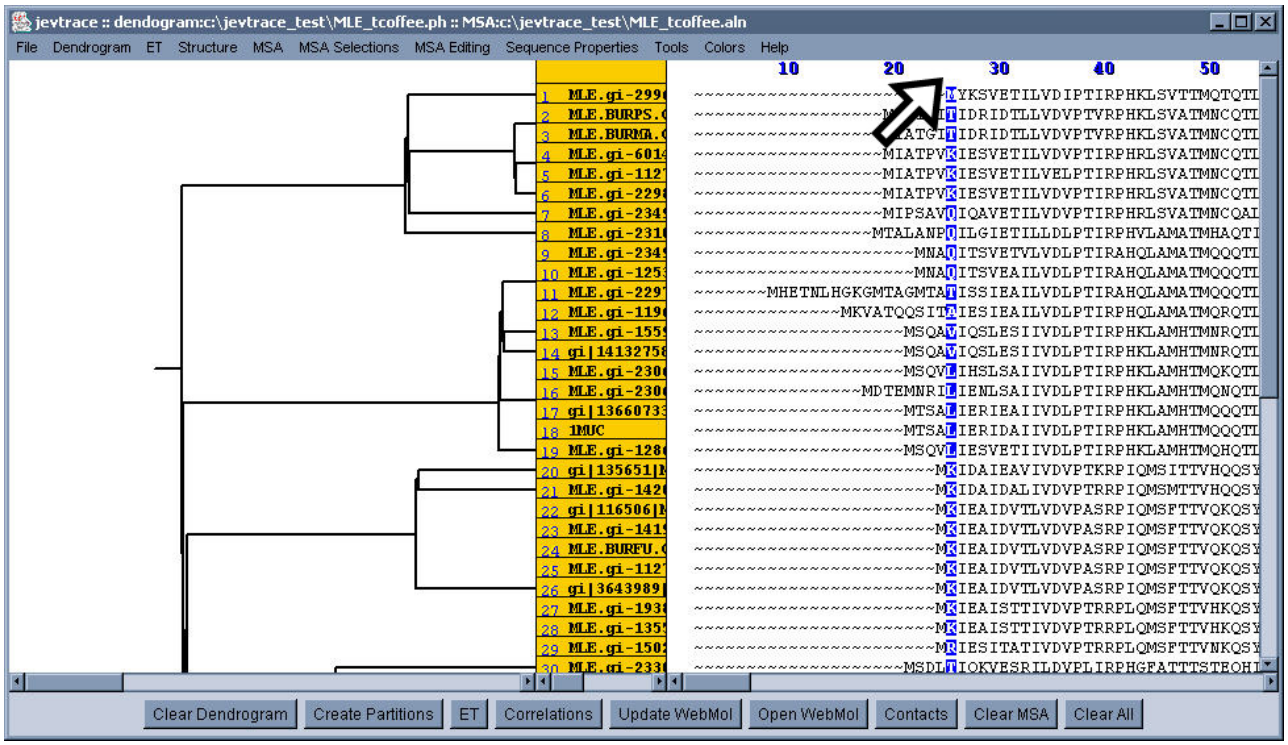


Figure 3
Mouse clicking above any column in the alignment, will select that column (or deselect if already selected).

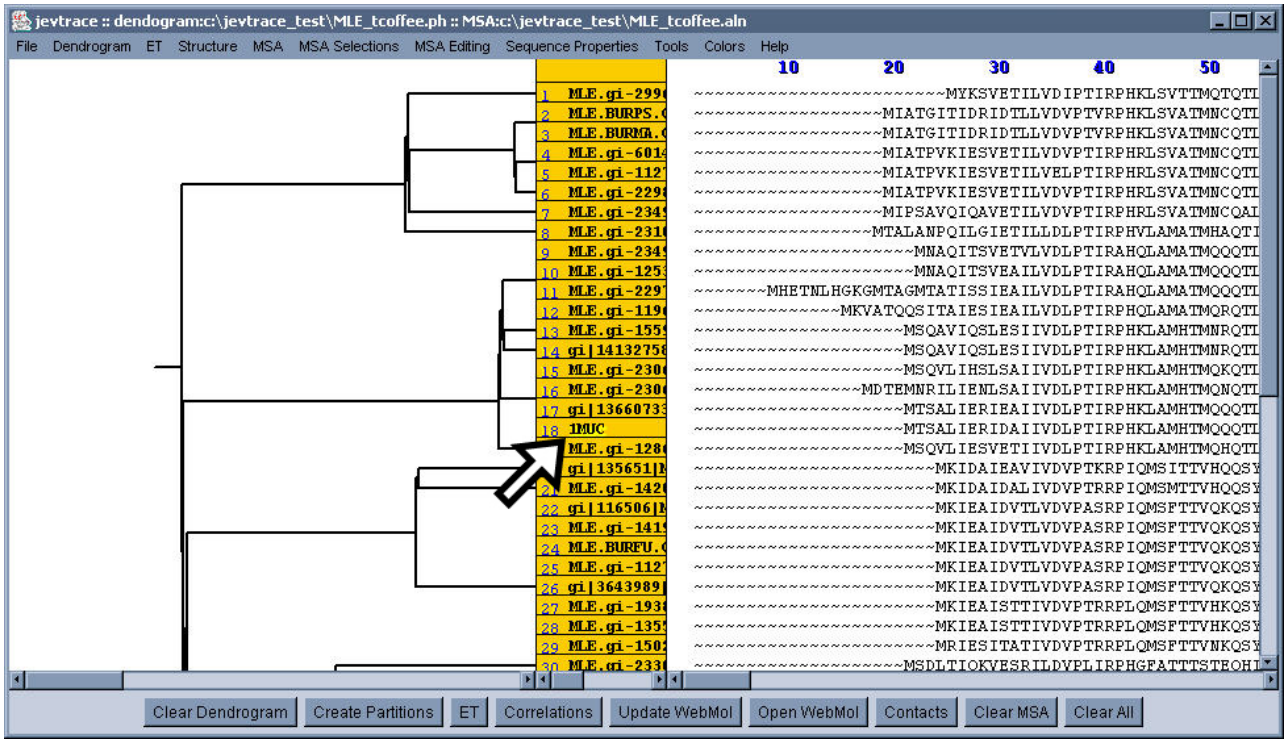


Figure 4
Clicking on any sequence name in the middle (orange) panel, will select that sequence column (or deselect if already selected).

4.2 Editing alignments

All of the alignment editing functions are in the '**MSA editing**' menu. Any result of an editing function can be reverted to the previous state (one step back in the history) by choosing the '**Revert to previous sequences**' option.

To remove a character or characters, simply select them in the alignment and choose the '**Remove character(s)**' option. Jevtrace will automatically pad the appropriate sequences with gaps at the end (to preserve the alignment format).

'**Remove selected columns**' and '**Remove not selected columns**' will perform the obvious operations based on the selected columns.

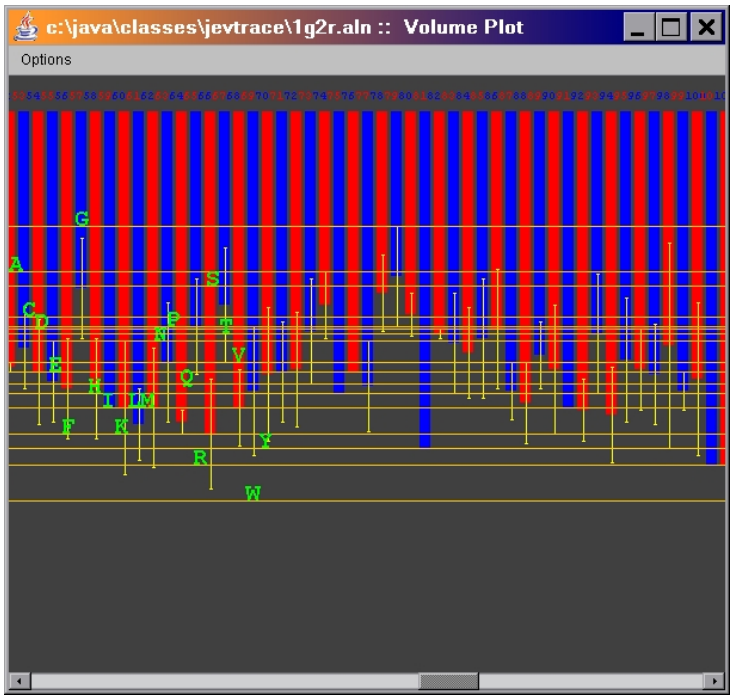
'**Remove selected sequences** and '**Remove not selected sequences** will perform the obvious operations based on the selected sequence names.

To add a gap character select a character, characters or columns and choose '**Add gap before'.** This will place gaps before any selections and will automatically pad the appropriate sequences with gaps at the end (to preserve the alignment format).

'**Left inclusive column cut**' and '**Right inclusive column cut**' will remove all columns to the left or to the right of the selected column (a single column selection is required).

4.3 Sequence properties

The '***Sequence Properties'*** menu provides analysis and graphical views of chemical and physical property analysis of the sequence data. The available properties include: molecular mass (from *Proteins*, Creighton, 2nd Edition), molecular volume (from *Proteins*, Creighton, 2nd Edition), electronegativity (from *Proteins*, Creighton, 2nd Edition), hydrogen bonding potential (counted as the total number of donor and acceptor groups), entropy (counted as the number of rotatable bonds), the oil/water partition coefficient (kcal/M per residue burial), and the Kyte-Doolittle hydrophobicity scale for transmembrane regions. The relative value for each amino acid is shown by the green letters and orange lines. The alternate blue and red coloring of positions is for contrast purposes only.



5. Making Multiple Alignments and Protein Structures Talk

All structure viewing in Jevtrace is designated to the Java structure viewer WebMol. Jevtrace and WebMol are distributed as an integrated package.

5.1 Loading and associating a structure in Jevtrace

To open a structure in Jevtrace the sequence of the structure has to be one of the sequences in the multiple alignment.

1. Select the sequence in the alignment corresponding to the sequence of the structure (Figure 4). The sequence should be identical in amino acids and length to the sequence of the PDB. Discrepancies will be reported in a warning dialog window.
2. Click the 'Open WebMol' button at the bottom of the Jevtrace window. A file dialog will prompt to choose a PDB format file.
3. The PDB file will be read by Jevtrace and the sequence of the PDB structure will be aligned with the selected sequence in the alignment. A warning dialog will report any discrepancies – in most cases the association will still be performed. If the file format is valid, WebMol will display the structure.

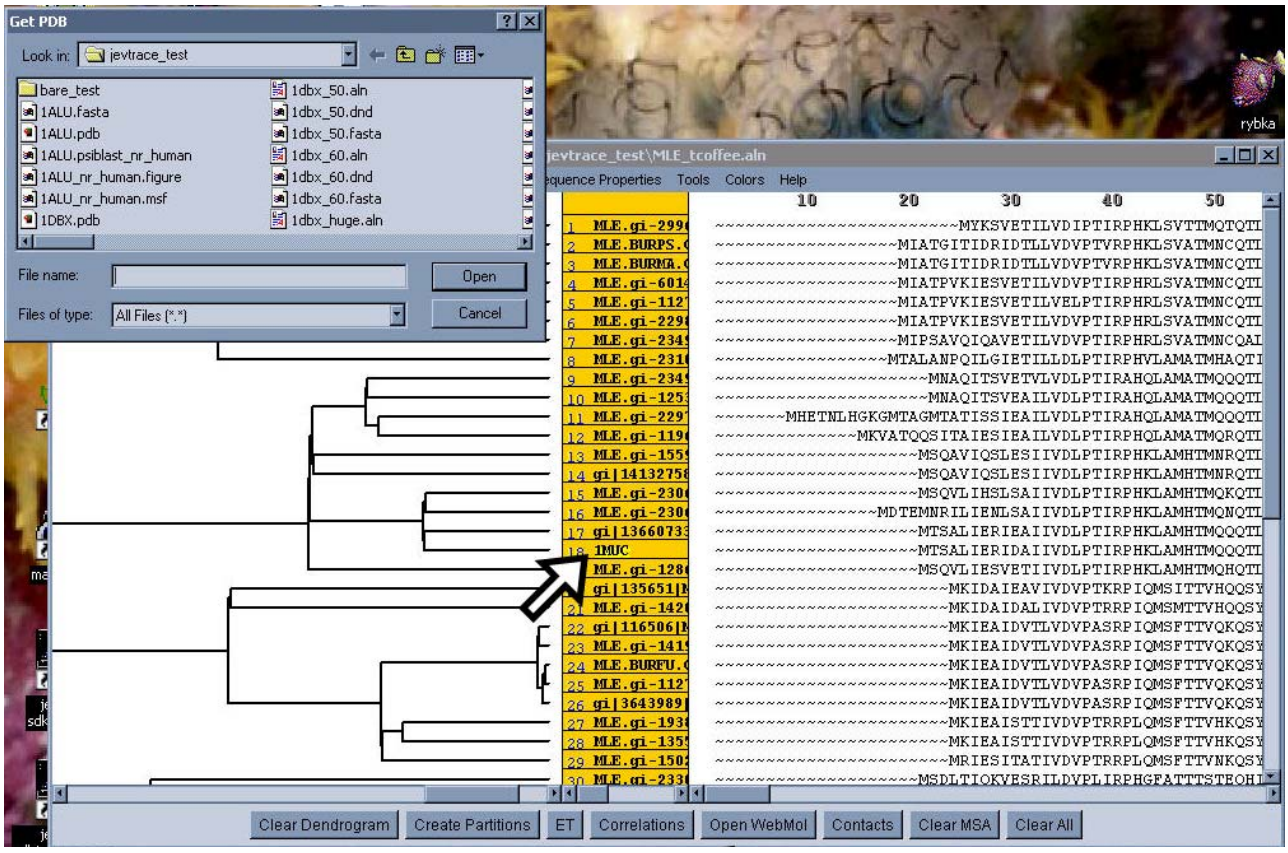


Figure 5
After selecting the sequence to be associated with a protein structure, mouse clicking on the 'Open WebMol' button will open a file dialog prompting for a PDB format file.

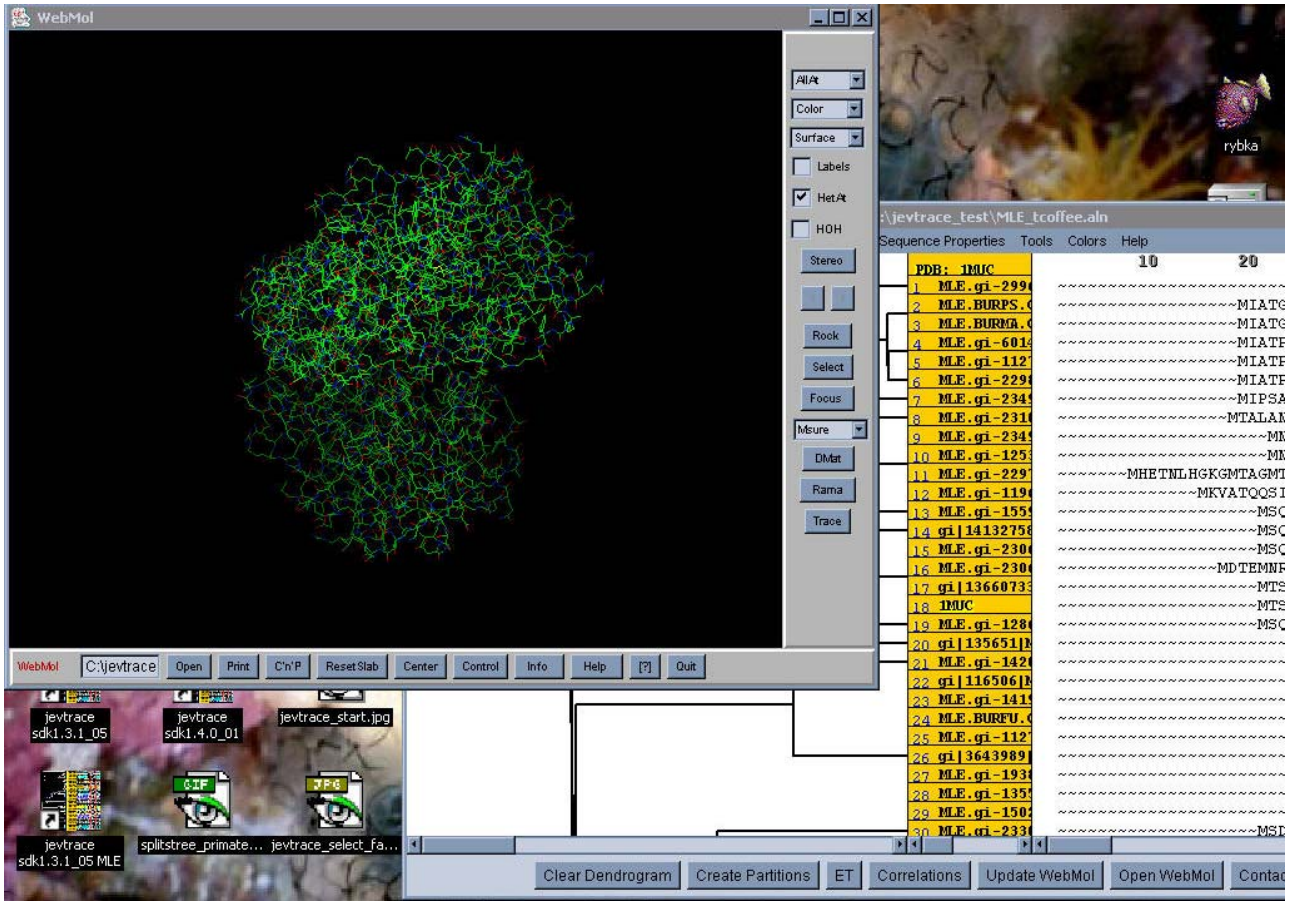


Figure 6

Selecting a valid PDB file will display the structure in WebMol. Jevtrace aligns the sequence of the PDB structure with the selected sequence in the alignment. This alignment is the basis for all subsequent sequence to structure mappings.

5.2 Displaying alignment selections on the structure

Any selections in the multiple alignment will be automatically displayed on the structure in WebMol. The **‘Structure’** menu has a ‘**Map Selected Sequence**’ option that will only map selections of that specific sequence.

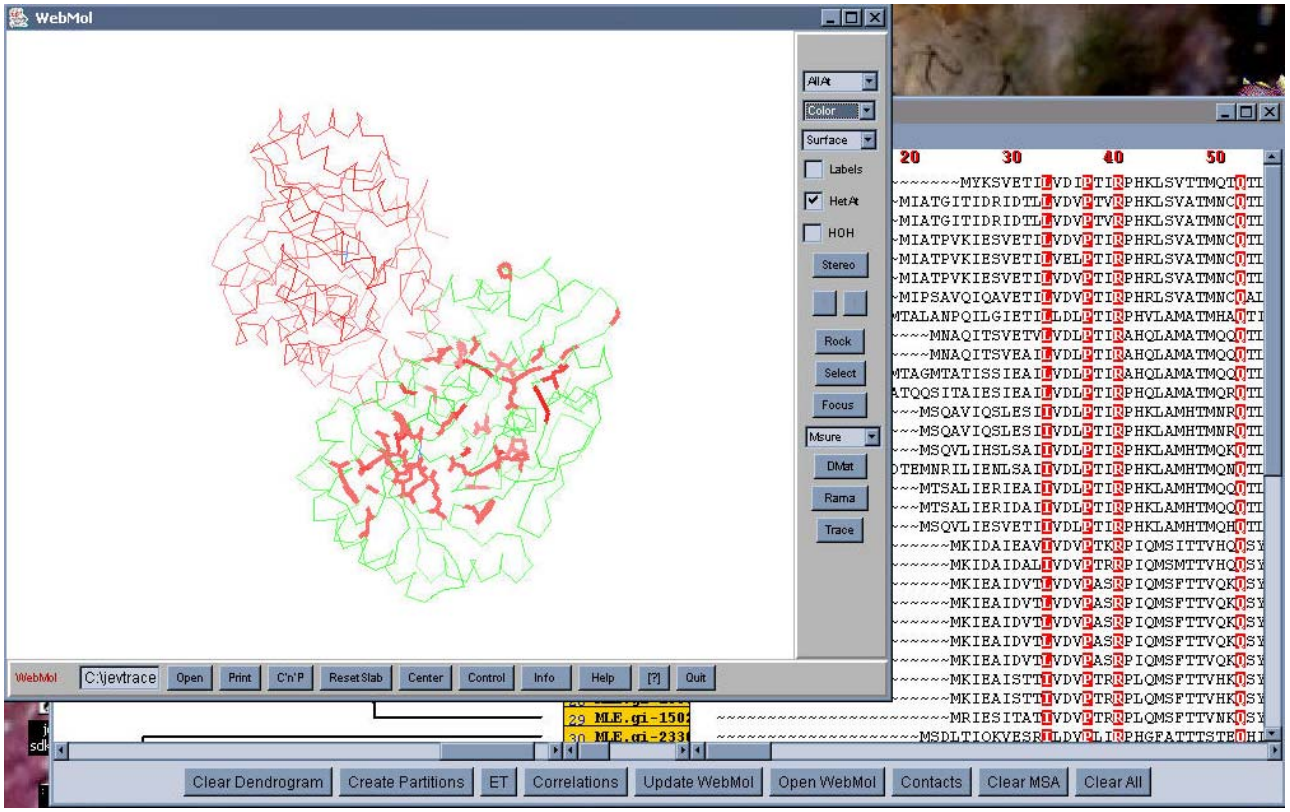


Figure 7
Selections in the sequence alignment are automatically displayed on the structure in WebMol. The background in WebMol was changed to white for higher contrast (in the color menu top right of the WebMol window).

5.3 Using multiple structures

During a single session, Jevtrace allows interaction with an unlimited number of protein structures; however, only one structure can be associated with the alignment at one time. In order to associate a new structure with the alignment, click on the previously selected sequence name (to deselect it) and select a new sequence name. Then click ‘Open WebMol’ and choose the new PDB file.

6. Making Trees and Multiple Alignments Talk

The tree functionality in Jevtrace allows use of branching and grouping information inherent in trees and phylogenetic trees.

6.1 Selecting subfamilies in the tree

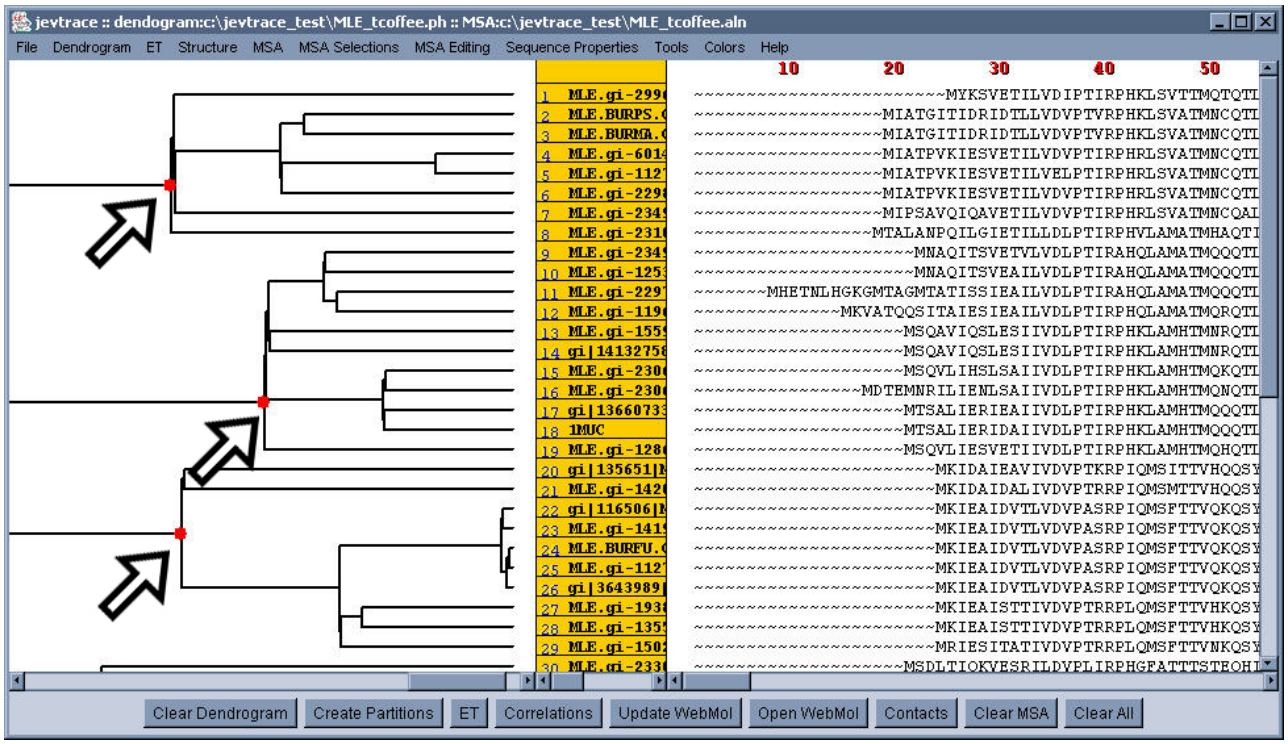


Figure 8
Selection of subfamilies in the tree.

To select a subfamily of sequences click on the middle (end) of any branch. Such selections are the basis for all advanced computations with Jevtrace, including calculation of subfamily sequence conservation, evolutionary correlations, tracing the ancestors of a node and the Evolutionary Trace.

When the distances between sequences or ancestral nodes are small it can be difficult to see which sequences belong to a branch – in this case click with the mouse while holding down the shift key. This will select all the nodes below that node.

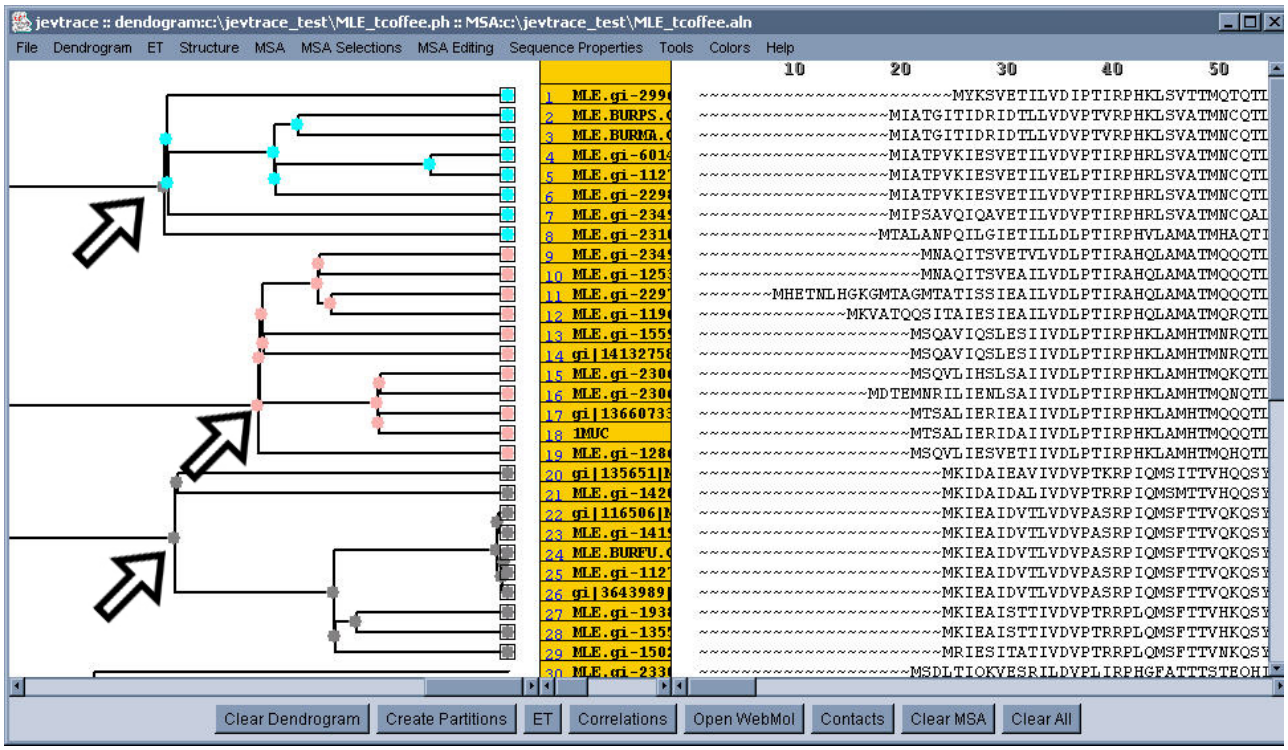


Figure 9
Using a mouse click while holding down the shift key will select all the subfamilies belonging to the originally selected node.

6.2. Jevtrace analysis using selected subfamilies

For example to analyze the evolutionary sequence correlations amongst the three subfamilies selected in **Figure 1**, simply click on the '**Correlations'** button.

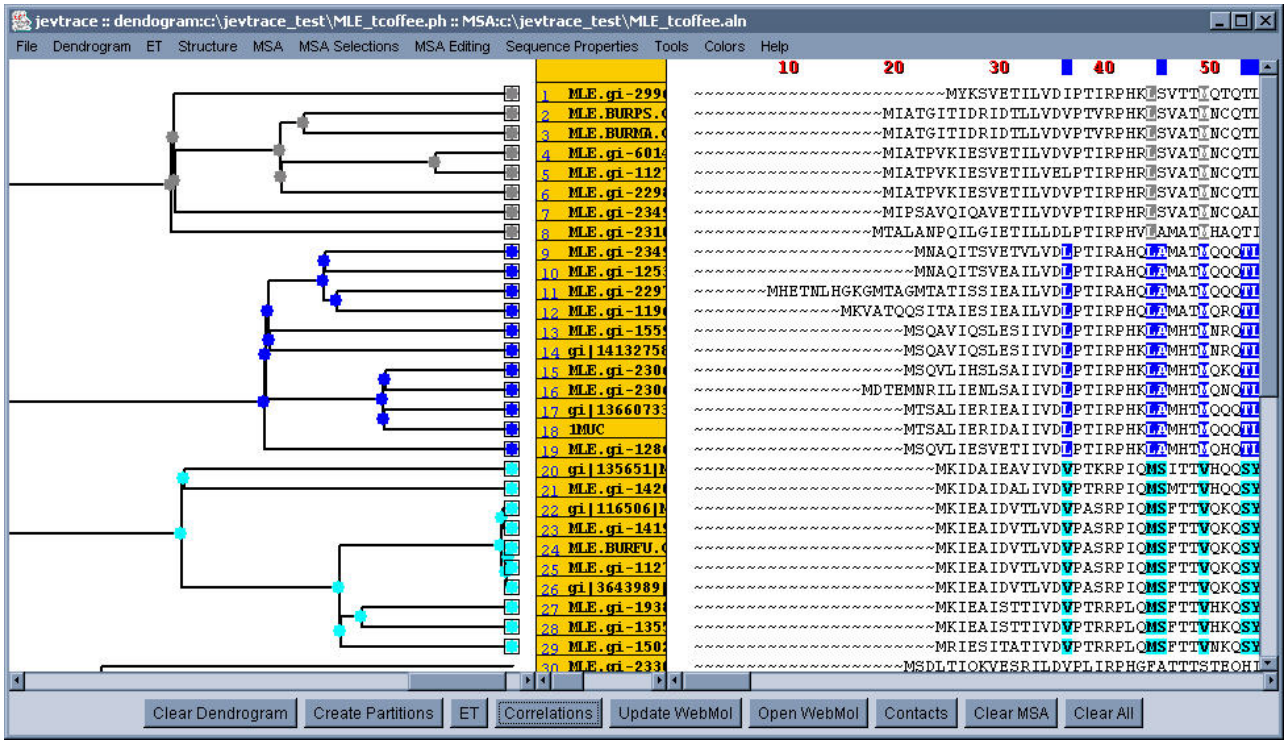


Figure 10
Results of evolutionary correlation analysis based on the subfamilies selected in Figure 8.

6.3. Jevtrace analysis using selected partitions – the Evolutionary Trace

The original Evolutionary Trace method relies on creating partitions of the phylogenetic tree. This procedure has been difficult to automate. Also the use of percent sequence identity may be misleading, since evolutionary models rely on more subtle measures of sequence similarity and distance. Jevtrace allows creation of up to eight partitions anywhere on the phylogenetic tree. The partitions can be created in any order – Jevtrace will sort them left to right when computing the Evolutionary Trace analysis ('**ET**' button).

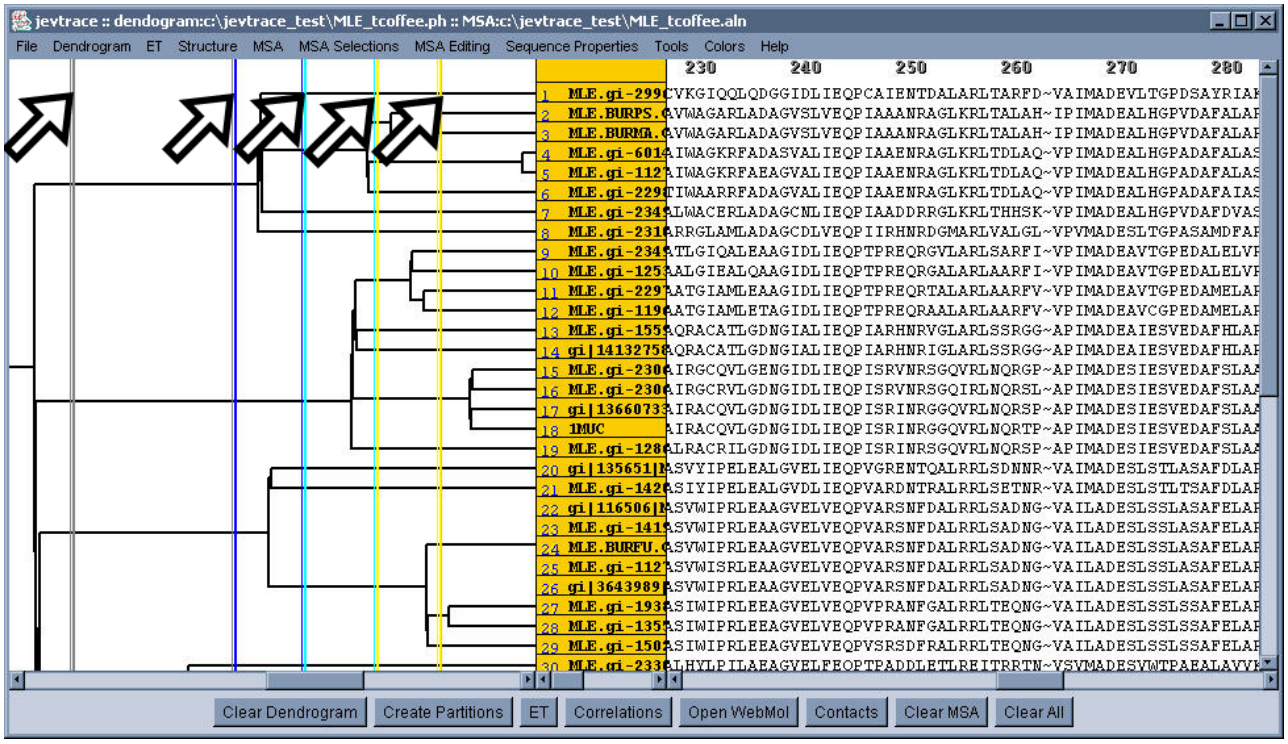


Figure 11
Creation of partitions of the phylogeny is enabled by clicking the '**Create Partitions**' button at the bottom of the Jevtrace window. Once enabled, clicking on the desired points in the phylogeny will produce a series of vertical colored lines.

Using the two functions in the '***Tree***' menu, the partitions can be saved to or read from a file. '**Write partitions from file**' will create a tab delimited text file with the relative position of each partition. '**Read partitions from file**' will load this data and display it on the phylogenetic tree.

To aid in the selection of partitions, the phylogenetic tree can be annotated with the average percent sequence identity of all sequences below each node. This is done using the '**Annotate with ungapped sequence identity**' function in the '***Tree'*** menu. Alternately, the branch distances from the original tree or tree file can be displayed using the '**Annotate branch length** in the '***Tree'*** menu.

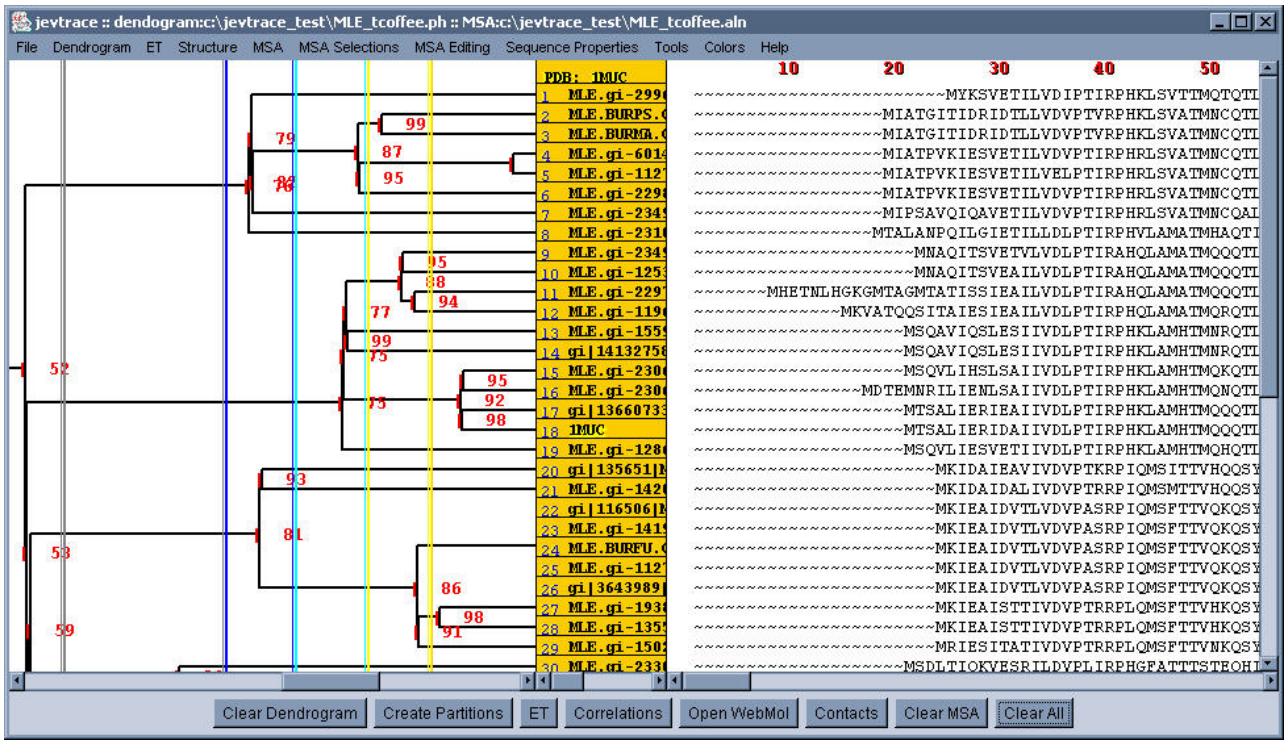


Figure 12
Displaying the average percent sequence identity for each node in the phylogenetic tree using the '**Annotate with ungapped sequence identity**' function in the 'Tree' menu.

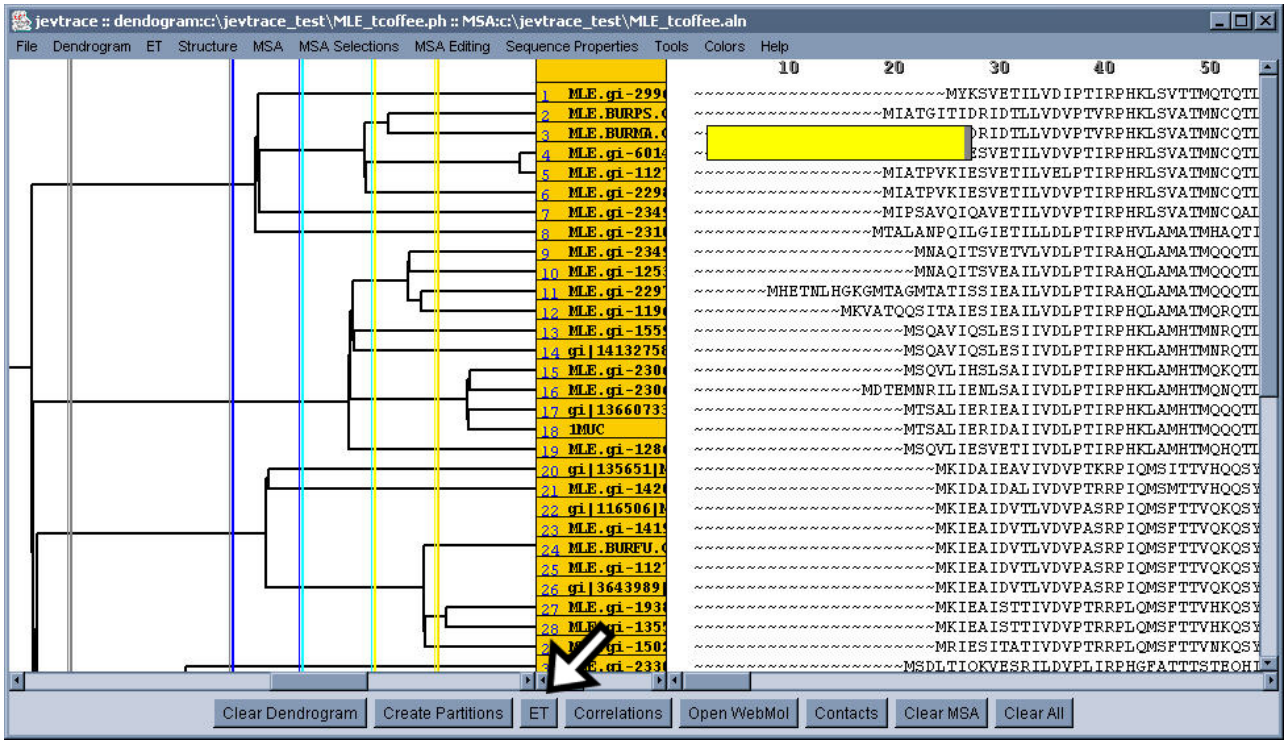


Figure 13
Clicking on the '**ET**' button at the bottom of the Jevtrace window, will compute the Evolutionary Trace residues for the selected phylogenetic partitions. Performance and approximate runtime will be monitored with a progress bar.

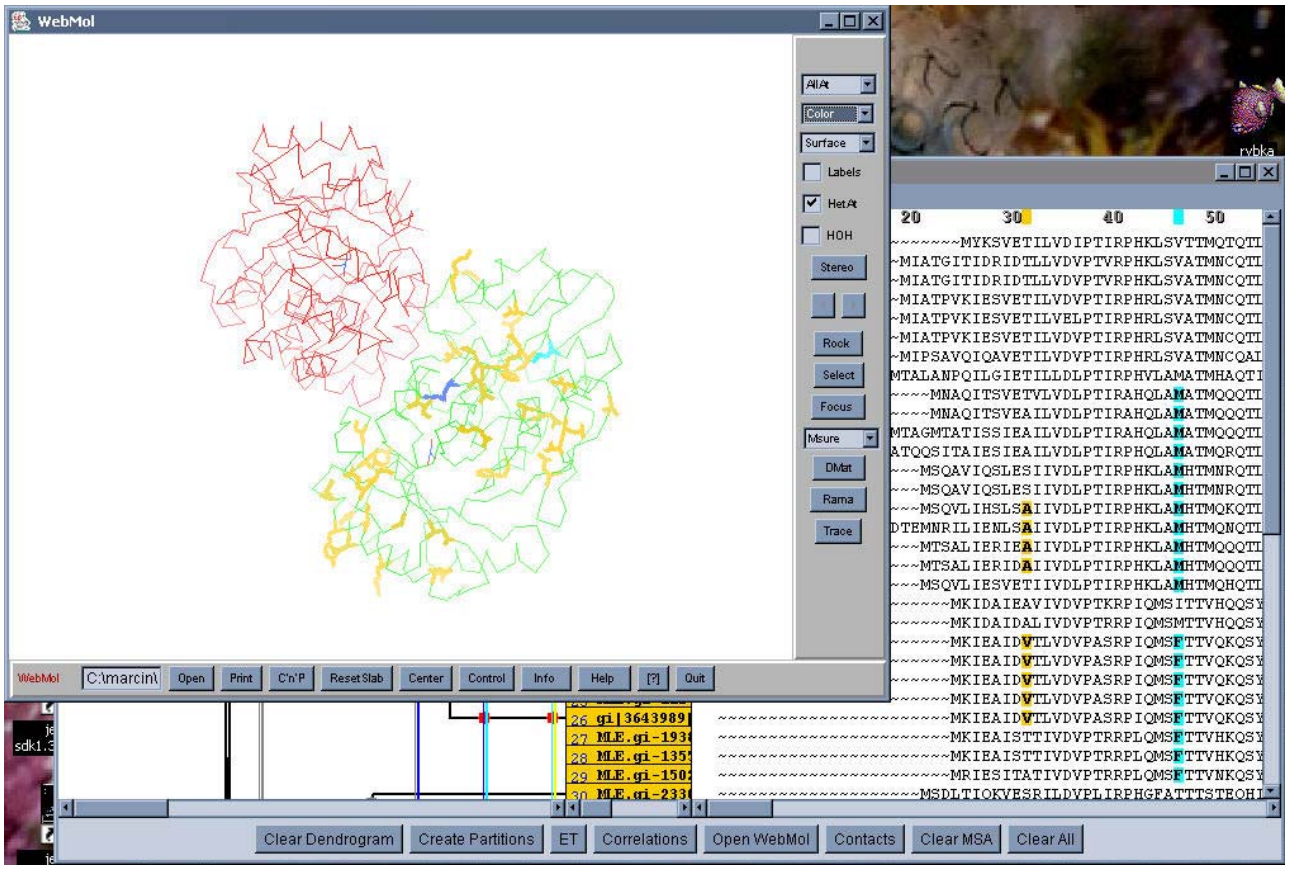


Figure 14

The Evolutionary Trace results will be shown as colored selections on the alignment, and as colored residues on the structure (if the structure has been previously opened in WebMol, see Section 5.1).

7. Dictionary of Menus and Features

A. Graphical Selections

In all cases, clicking on a graphical object selects and clicking again deselects.

TREE

single node selection - mouse click on node in tree.

subfamily selection for analysis - hold ctrl key and mouse click on the parent node of a subclade.

graphical highlighting of a subfamily – hold shift key and mouse click on a node.

NAME CANVAS

sequence name - mouse click on a name. A selected name is required for opening a protein structure. Multiple names can be selected for multiple alignment editing and analysis.

ALIGNMENT CANVAS

single amino acid - mouse click over amino acid in multiple alignment.

alignment column selection - mouse click above the column (in the region of sequence numbering).

B. Buttons

Clear Tree

Create Partitions - enables creation of partitions of the phylogenetic tree. Partitions are created by clicking on the tree canvas. Up to eight partitions can be created in any order (they will sorted by Jevtrace from left to right).

ET - based on user selected partitions (defined above in '**Create Partitions**') or a set of selected nodes, performs the Evolutionary Trace computation. The results will be shown on the sequence alignment (and structure if it is associated). The score for each column is shown above the sequence alignment.

Correlations - based on a set of selected nodes, performs a evolutionary correlation computation The results will be shown on the sequence alignment (and structure if it is associated). The score for each column is shown above the sequence alignment.

Clear Names

Open WebMol – requires a selected sequence in the alignment. The sequence must be identical (amino acids and length) to the sequence of the structure. User is prompted for a PDB file - Jevtrace aligns the PDB sequence to the selected sequence. If the alignment is successful displays the structure in WebMol. This operation is required for any structure-based computations.

Contacts - requires an open protein structure. For any selections in the MSA (single seq or entire row), computes residue-residue contacts of those positions. An all atom comparison is made, using 6A as the contact cutoff.

Ligand contacts – calculates the residues contacting the ligand (if the PDB file has heteroatom records). An all atom comparison is made, using 6A asthe contact cutoff.

Clear MSA

Clear All - clears all selections in the MSA and tree canvases (but not the Name canvas in order to avoid breaking and sequence-PDB associations).

C. Menus

a. File

Open Tree and MSA - open paired data consisting of tree and multiple sequence alignment. A dialog first prompts for the tree file, and then for the alignment.

Open MSA

Open Tree

Open Session File - opens a saved session file and recreates any associated PDB, sequence, tree and name selections.

Save Session - saves a session.

Quit - quits the program.

b. Tree

Annotate with ungapped percent sequence identity - annotates the tree with the percent sequence identity over all non-gapped residues for each subclade.

Annotate with branch length- annotates the tree with the branch lengths.

Show consensus for selected nodes – displays the computed consensus sequences for any node(s) selected in the tree.

Color Positions Conserved Below Node - finds all positions in the selected subclades that are conserved within those subclades and different between them.

Trace parents of a node – for a selected node traces the sequence conservation in up to eight ancestors (parents).

Read tree selections from file – reads a text file with encoding all of the tree graphical selections.

Write tree selections to file – writes a text file with encoding all of the tree graphical selections.

Average percent sequence identity in partition(s) – for the selected partitions computes and displays the average percent sequence identity of all the nodes in that partition.

Min/Max sequence identity in partition(s) – for the selected partitions computes and displays the minimum and maximum percent sequence identity of all the nodes in that partition.

Export MSA graphics (JPEG*)* – exports the alignment with colored selections to a JPEG file.

c. ET

Use MSA selection – during ET or correlation analysis, uses only the selected alignment columns for computations.

Convert to absolute sequence variability – after ET or correlation analysis, normalizes the column score by the absolute sequence variability at that position.

Show correlation score – after ET or correlation analysis, shows the pairwise subfamily score.

Show lowest rank score - after ET analysis, shows the lowest phylogenetic rank for that column.

Convert score to column scoring – after ET or correlation analysis, converts the column scoring to coloring of the entire column (for example this is necessary to output coloring compatible to Chimera).

Remove score – after ET or correlation analysis, removes the column scoring altogether.

d. Structure

Open WebMol – prompts for a PDB file through a file dialog. Loads the selected PDB file and aligns the PDB sequence to the selected sequence in the alignment. If the alignment is successful, associates the structure with the alignment and displays it in WebMol.

WebMol buried/exposed routine – calculates the residue solvent accessibility in WebMol. Buried residues are represented by thin lines and exposed residues as thick lines.

Note: the following two options can be combined

Map Selected Sequence - maps only positions for which this sequence is selected (excluding columns)

Map Selected Columns - maps only selected columns

e. MSA

Color Conserved Positions - colors absolutely conserved positions **red**.

Color variable positions – colors not absolutely conserved positions **blue**.

Color gapped positions – colors gapped positions **green**.

Color positions by percent sequence identity – colors all the positions in the alignment using the Jevtrace color scheme (green to red, where red is absolutely conserved).
> 99 = **red**
> 89 = **magenta**
> 79 =**pink**
> 69 = **orange**
> 59 = **yellow**
> 49 = **cyan**
> 39 = **blue**
> 29 = **gray**
<= 29 = **green**

Make sequences by pairwise identity – a dialog prompts for the percent sequence identity cutoff. For any pair of sequences with identity above the cutoff, the second sequence will be removed.

Filter sequences by minimum ungapped length – a dialog prompts for the minimum ungapped length cutoff. All sequences below the minimum are discarded.

Define similarity relative to tree selection – for the selected sequence, computes the unique residues relative to other sequences.

Define specificity relative to tree selection (discrete) – for the selected sequence, computes the unique residues relative to other sequences.

Define specificity relative to all – for the selected sequence, computes the unique residues relative to other sequences.

Write in Fasta format – outputs the current alignment to a Fasta format file.

Write alignment in ClustalW format (.aln) – outputs the current alignment to a Clustalw .aln format file.

Write alignment in Nexus format (.nexus) – outputs the current alignment to a MSF format file.

Write alignment in Phylip format (.phylip) – outputs the current alignment to a MSF format file.

Export MSA graphics (JPEG*)* – exports the alignment with colored selections to a JPEG file.

f. MSA selections

Undo last sequence selection – reverts to the last sequence selections in the history.

Save sequence selections to file (.seqsel) – saves all the sequence selections to a text file in SeqSel format.

Load sequence selections from file (.seqsel) – reads sequence selections from a text file in SeqSel format.

Save coloring (old versions of Chimera) – provided for backwards compatibility with older versions of chimera (before August 2003).

h. Sequence Properties

Mass Plot

Volume Plot

Electronegativity plot

Hydrogen bond plot

Entropy plot

Oil/water partition coefficient

Kyte-Doolittle hydrophobicity scale

g. MSA editing

Undo last edit – reverts to the sequences one step back in the history.

Insert gap – adds a gap for every selection in the alignment.

Remove all gap columns – removes all columns that are all gaps.

Remove character(s) – removes any columns or characters selected in the alignment.

Remove selected column(s)

Remove not selected column(s)

Remove selected sequence(s)

Remove not selected sequence(s)

Cut columns (left inclusive) – for a single selected column, removes all columns to the left of the selection, including the selection.

Cut columns (right inclusive) – - for a single selected column, removes all columns to the right of the selection, including the selection.

Export MSA graphics (JPEG*)* – exports the alignment with colored selections to a JPEG file.

i. Sequence Names

Name string search – searches the sequence names with a query string, selects matches.

Append unique number prefix – appends an integer prefix to every name.

Strip numbers to make sequence names unique – strips any existing integer sequence name prefix.

Convert names to species names – extracts the species name and designates them as the new sequence names (following the SwissProt convention e.g. LCA_HUMAN or LCAA_HORSE).

Export sequence names graphics (JPEG*)* – exports the alignment with colored selections to a JPEG file.

j. Tools

Write sequence distances to file – outputs the percent sequence identity for each unique pair of sequences to a tab delimited text file. The file consists of the name of the first sequence, name of second sequence, and the percent sequence identity to two significant digits.

Write pairwise correlation data – will generate a tab delimited text file encoding the pairwise sequence correlations based on the results graphically displayed on the alignment. Currently only a binary (0 or 1) correlation scale is used, and only residue pairs with non zero correlation data are output.

Export all graphics (JPEG) – creates one JPEG graphic of the phylogenetic tree, sequence names and sequence alignment.

k. Colors

Options in this menu allow selection of the active Jevtrace color. Red and green are special colors. It is important to pay attention to color conflicts when performing multiple nested analysis (e.g. color by percent identity, then ET, then correlation).

Color control – brings up a dialog which allows control of which colors are displayed on the alignment and structure.

Use smooth color scale – Jevtrace was originally designed with a discrete colored scale due to limitations in the structure graphics. There is now the option to generate smooth scale coloring, and this option allows to toggle (translate) between the two scales.

Coloring schemes:

Amino acid groups – hydrophobic/orange = AFILMVWY, hydrophilic = NQST, basic/blue = HKR, acidic/red = DE, cysteine/yellow = C, proline/magenta = P, glycine/pink = G

Amino acid groups + gaps - hydrophobic/orange = AFILMVWY, hydrophilic = NQST, basic/blue = HKR, acidic/red = DE, cysteine/yellow = C, proline/magenta = P, glycine/pink = G, gaps/black = - or ~ or .

Phosphorylatable (S,T,Y) – STY = green

Alignment quality (Blosum62) – BelVu style alignment quality coloring based on three similarity cutoffs.

I. About

About - Lists the Jevtrace version and system settings.